# Repertory Grid Technique : A Tool for Exploring Classroom-Based Assessment?



| | |
|---|---|
| journal or publication title | Kwansei Gakuin University Humanities Review |
| volume | 25 |
| page range | 13-28 |
| year | 2021-02-18 |
| URL | http://hdl.handle.net/10236/00029286 |

*Kwansei Gakuin University*
*Humanities Review*
Vol. 25, 2020
Nishinomiya, Japan

# Repertory Grid Technique: A Tool for Exploring Classroom-Based Assessment?

Myles GROGAN*

## I. Introduction

Under a classical approach to testing, a clearly defined construct is measured while keeping as many other variables as possible stable. Such psychometric measurement principles, essential to large-scale testing, may not be appropriate for a classroom setting, however (Turner, 2012), particularly in non-major subjects, such as English as a foreign language (EFL) in a university setting. In an age where universities are demanding greater accountability of teachers (Hadley, 2015) and where assessment literacy is increasingly on the research agenda (Coombe et al., 2020), understanding what a classroom grade means is essential. However, in many places like Japan, given the strongly localized approaches to assessment from different schools and even different teachers within the schools, generalization may be difficult (Hill, 2017). In particular, the kind of information that a grade carries is hard to define, even though the grades awarded for compulsory classes commonly form part of a student's grade-point average. As such, the grading process is consequential. This paper therefore investigates an alternative approach to understanding the classroom grade, by asking teachers to consider what may be inferred from the scores they award, using a Personal Construct Psychology tool known as the Repertory Grid (Denicolo et al., 2016; Hadley, 2017).

Much of the guidance for large-scale testing, such as that offered by the International Language Testing Association (ILTA, 2007), may be neither desirable nor practical in many classroom settings (Gipps, 1994; Smith, 2003). Recent approaches to classroom-based assessment have emphasised the use of formative assessment (Brookhart & McMillan, 2019; Heritage & Harrison, 2020). There has

* Instructor of English as a Foreign Language, School of Science and Technology, Kwansei Gakuin University

also been an increase in awareness of the impact the sociocultural turn (Lantolf & Poehner, 2014; Swain et al., 2015) and of the impact of the community on learning. Increasing attention is also being given to the experiences of students before arriving in an assessment situation. The different "opportunities to learn" impact both student performance on assessment (American Educational Research Association et al., 2014), as well as broader participation within a classroom community. The teacher has become responsible for both nurturing learning and for evaluating the learning, roles which may be seen to be in conflict (Bishop, 1992; Cheng & Fox, 2017).

Research into classroom-based assessment has recognized that teachers use the main constructs of language content as well as broader academic enablers (Sun & Cheng, 2014). Formal work on teacher and student assessment literacy continues, often using questionnaires or survey instruments (Brown & Hirschfeld, 2008; Harding & Kremmel, 2016). However, alternative approaches to assessment research, such as narrative inquiry using student pictures, have uncovered areas missed by questionnaires (Brown & Wang, 2013). More direct and open techniques may help to uncover aspects of the teacher or student experience in a way that is more in touch with the experience of the classroom (Hadley, 2017).

The Repertory Grid (RGT) is a tool from Personal Construct Psychology (Kelly, 1992). It has been used for simple classroom research (Block, 1997), and has played a small but significant role in research into teacher thinking (Borg, 2015). Because it focuses on the subjective, it is particularly suitable for investigating perceptions of teachers. A simple outline of the process for use in applied linguistics is available in Hadley (2017), with more detailed general background to the process available from Jankowicz (2004) or Denicolo *et al* (2016). The process can be used as part of a broader qualitative data analysis approach (Bazeley, 2013; Miles et al., 2014) and can be used in conjunction with specific approaches such as Grounded Theory (Bryant, 2017) or Thematic Analysis (Guest et al., 2011).

The technique resembles a coding process conducted with a participant, in which an interviewer asks a participant to compare a small number of "elements," so that the researcher and participant can more fully understand how the participant construes them. The constructs are made to be bi-polar. This means that the participant actively creates a discrete meaning for each end of the pole (e.g. "black" and "white") rather than simply negating something ("black" versus "not black"), allowing a more thorough discussion of the elements being construed. In the current inquiry, the interviewer asks a teacher to consider students receiving different grades, and to create scales of constructs on which to compare the students. This allows the researcher to see some of the processes that may underlie the grading

process, while responding to recent calls in the literature to be "ecologically sensitive" (Fulcher, 2010, p.2).

As the participant rates each element on a scale for the construct, quantitative data is available in addition to the qualitative. This allows researchers to compare the constructs to a desired construct, such as a grade, using a process for comparing similarity (Honey, 1979), allowing teachers to potentially reverse engineer the main grading construct (Davidson & Lynch, 2002). The quantitative data also allows for the use of principal component analysis, making the technique a kind of inherently mixed method (Bazeley, 2018). This makes it particularly appealing from a research point of view.

In part, this project represented a kind exploratory practice (Allwright, 2005; Hanks, 2017) on the part of the researcher, in that it was partly to better understand repertory grids as a possible research tool, and the practices that go into classroom grading. Additionally, it was done with the cooperation of a teacher with similar interests. This enquiry therefore sets out to try the technique and to judge the strengths and weaknesses of repertory grids as a research tool that may help both classroom and assessment research, as a prelude to further inquiry.

Setting for the pilot study

The educational institution in which this study is set may be described as a mid -level private Japanese university. The class that was discussed with the participating teacher was an instance of the required speaking and listening course for all first-year students, as part of a general education program. For the purposes of this paper, the course will be called *English A*. The curriculum is the same for all faculties, regardless of field, and is overseen by a single faculty, rather than a "third space" such as a language centre (Hadley, 2015). According to institutional materials, the goals of the *English A* course are for students "to understand spoken English on general, daily topics without much difficulty" and to "be able to express their ideas orally with basic words and simple sentences." Although the guidance in the curriculum suggests some pedagogic activities, specific learning outcomes are not present in the curriculum. Students are placed in classes within each faculty

**Table 1** ***Grade Quota and Breakdown***

| Grade Quota | Grade Breakdown |
|---|---|
| S: 10-20%<br>A or B: 60-80%<br>C or fail: 0-20 | 20% − Common test material<br>10% − Short tests<br>20% − Midterm test<br>20% − Final test<br>30% − In-class activity |

based on the result of a general commercial placement instrument. The majority of students are at the CEFR A1 level, although strengths and weaknesses vary considerably.

Roughly 80 teachers teach this course, and each one must use selections of required material in activities to create a defensible grade. The grading uses a quota-based system, shown in Table 1. The number of students who may receive a specific grade is fixed. For example, in the class used in this study, 10-20 percent of students must receive the highest grade (the "S" grade). The components of their scores are also shown in Table 1, though teachers submit their whole grade, and not the components of each score. Teachers are free to create a system for assigning these numbers that fits their own processes, although they may have to defend the grade in the event of an inquiry.

## II. Research Question

Using a qualitative data analysis approach (Bazeley, 2013; Miles et al., 2014), this paper attempts to explore the suitability of Repertory Grid Technique using the following guiding question as a central theme (Creswell, 2014):

> *What qualities does the teacher observe or infer in students at each of the grade levels in the English A class?*

## III. Method

Participant and class

A single teacher collaborated with the researcher, consenting to the use of the interview for publication and for the video recording of the interview. The participant was a former colleague of the researcher, representing a convenience sampling that was suitable for the purpose of trialling a new technique. The teacher has been in Japan for over 10 years, and was a contract teacher under the auspices of the faculty managing *English A* program. As with all contract teachers, he has an MA in the field of TESOL and Applied Linguistics. For the procedure, he chose to talk about an intermediate class, meeting midweek on the first period. It was a large class, with 38 members.

The teacher designed the class around speaking instruction. The teacher explicitly stated that he thought listening would be acquired as part of the speaking process, and therefore focused on assessing communicative speaking tasks in formal assessment. Listening tasks from the textbook were used to the extent that they

supported uptake of the theme for use in speaking activities. Similarly, language-form or vocabulary tasks from the text were used to prime the students for completion of the main assessed communication tasks, rather than being discretely tested in their own right.

The main assessed learning goals of the teacher's instance of the course were for students to give presentations on simple topics and to participate in discussions. Comprehension of the unit topic was demonstrated through presentations. After the presentation, students were given two weeks to prepare for a graded class discussion on the same topic. The discussions (described as "like a panel discussion") allowed for a focus on more interactive and daily communication skills, with the teacher giving examples such as asking for opinions or getting clarification as elements of grading. These tasks combined, given twice a semester, made up 50% of the grade. In addition, homework was given and graded using a cumulative system as "completed" or "not completed" (15% of the grade). Participation in the class was also graded (15% of the total grade for the class), promoting uptake of positive behaviours. Finally, a compulsory online vocabulary course that all students in the program had to complete made up the final 20% of the grade.

Interview Procedure

The interview was given in the second semester, after the first round of grades for the class had been delivered. The interview took about 80 minutes, and was modelled closely on guidance given in Jankowics (2004). Biographical information was gathered, followed by a description of his class in general terms (given above). Using examples of constructs from personality, taken from Jankowics (2004), guidance was prepared for the participant on how to make constructs. This was in the form of a visual aid in the opening stages of the interview and served as a reminder throughout the interview.

For this procedure, the "elements" were students from the class described who received different grades in their first semester (S, A, B, C, or D). The participant chose eight students. A description of each of these students was elicited. The teacher noted that, because most students had achieved a near perfect score on the compulsory vocabulary component of the grade, many students got a high score, and several had had to be downgraded prior to grade submission. This information

**Table 2** ***Grades and Pseudonyms of Students***

| Name | Masako | Ryo | Koh | Makiko | Chieko | Osamu | Akiko | Yukiko |
|---|---|---|---|---|---|---|---|---|
| Final grade | S | S | A | A | B | C | C | D |
| *Raw grade* | *S* | *S* | *S* | *S* | *B* | *C* | *D* | *D* |

can be seen in Table 2 (below), with the final grade on the top row, and what the participant called the "true" grade below. He reported only a single B grade in his class, representing a feature of the grading procedure rather than sampling.

A sheet for recording the specifics of the interview was prepared (see Appendix 1), and the interviewer made field notes during the interview process. In this process, the participant was asked to consider three of the eight students. He was asked how one student differed from the other two. To the extent possible, this difference was probed and framed in a bi-polar way to avoid simple negation. As an imaginary example, two students may often ask questions in class. Rather than having the opposite as "Does not ask questions," a preferred (positive) option might be "seems to avoid attracting any teacher attention." This is called *triadic elicitation.*

Following elicitation and concept checking, the members of the subsample were used as opposite poles of an integer scale (1-5), based on the construct, with which to compare the other members of the sample. The remaining five students were rated on the construct created by the participant, with any issues or difficulties, such as members who do not fit the construct, addressed through negotiation. The process was then repeated with a different subset of elements. This comparison produced a bi-polar scale, called a "construct".

After an initial set of constructs had been elicited, the participant was invited to create any constructs that they thought may be relevant but that had not yet arisen (Denicolo et al., 2016). Finally, the participant rated the students on their English speaking, listening, and overall English language ability. In RGT literature, these are *supplied constructs* (Jankowicz, 2004), in that they come from the interviewer, rather than the participant.

In total, seven original constructs were derived. Combined with the supplied constructs, the result was 10 rated scales. The session concluded with a review of the students' information as a concept check, to confirm the placement of students with the participant.

Analytic Process

The interview was transcribed, and summaries made both of the teacher's biographical data and the student profiles. This enables familiarization with the data (Guest et al., 2011). Memos were added on points of interest, including notes on the kind of constructs derived, such as the extent to which they were behaviour or skill-based, observational or attributive, core to the idea of grading or perhaps more peripheral to the grading process.

The numeric data was analysed using the WebGrid Plus online system (*WebGrid Plus,* 2017), using two main features. The dendrogram shows the numeric

relationship between individual constructs and then individual elements. A mapping function shows a Principal Component Analysis (PCA) shows relationships between the constructs visually, in the form of a graph, as well as coming up with a number of "components" to explain the variance.

## IV. Results

The constructs and scores elicited for each element are shown in Table 3. The interviewer and the participant collaborated to concept check these constructs as far as possible during the interview. The interpretation presented here is therefore co-constructed, following repeated passes at the data. In addition to the grades, a "Similarity score" is presented on the left-hand side of the table (Honey, 1979). This score looks at how a construct is similar to the final grade, with the S grade as a "1," and the D grade as a "5." The highest similarity in this case is the ability to handle the material (C01), at just over 81 percent, with the supplied constructs of language ability following closely. C05, with the lowest score, is the least likely to contribute to a general model of grading and grade performance.

In the opening of the interview, the participant noted that students made groups with their own gender, with the exception of one student equally comfortable working with either gender (Makiko). While the participant tried to create a construct around this, he was unable to do so in a bi-polar way. The topic was of interest, but did not seem to fit the research process. The rest of the interview, however, saw a detailed exploration of how the participant perceived those students that he taught in each different grade category.

The post-interview analysis began with the interviewer trying to categorize the kind of constructs obtained in the interview. Three constructs created by the participant were judged to be evaluative (C01, C03, C07), meaning they required judgement on the part of the participant. These seemed to relate to core aspects of the grade, although "Mentally prepared/Only physically present" (C07) perhaps spoke more to disposition. In contrast, C04 seemed to speak to a student's value system as perceived by the participant, making it an attributive construct. The participant suggested valuing the topics might be connected to the overall grade to some extent, perhaps as a moderating factor. Finally, C02, C05, and C06 were considered to relate more to observable student behaviour. The participant and the interviewer discussed C06 and C07 at some length, in order to distinguish them, noting that some students had materials such as textbooks or pens, but failed to use them. Although these are different categories of constructs, the participant deemed both these constructs to be associated with the grade outcome. Finally, C05 seemed to be peripheral to the grade, with less emphasis from the participant. He made it

**Table 3 Summary of Scores Elicited from Participant**

| Construct | Explicit pole | Masako | Ryo | Koh | Makiko | Chieko | Osamu | Akiko | Yukiko | Implicit pole | Similarity score |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Final score | S | S | A | A | B | C | C | D | Final score | |
| | Grade | 1 | 1 | 2 | 2 | 3 | 4 | 4 | 5 | Grade | |
| C01 | Able to handle material | 1 | 1 | 2 | 2 | 3 | 4 | 5 | 3 | Less able to handle material | 81.25 |
| C02 | Leads group | 1 | 1 | 2 | 3 | 3 | 3 | 5 | 3 | Follows groups / doesn't initiate / dependent on lead | 68.75 |
| C03 | Stays on topic | 1 | 4 | 1 | 1 | 2 | 5 | 5 | 3 | Diverges / gets distracted | 37.5 |
| C04 | Seemed to value the topic in context | 1 | 4 | 1 | 1 | 2 | 5 | 5 | 3 | Seemed disinterested in the topic | 37.5 |
| C05 | More chatty | 1 | 1 | 2 | 5 | 1 | 1 | 1 | 4 | Quieter | 25 |
| C06 | Physically prepared | 1 | 1 | 1 | 1 | 2 | 3 | 5 | 3 | Not physically prepared (no book, pens, etc) | 56.25 |
| C07 | Mentally prepared | 1 | 3 | 1 | 1 | 2 | 5 | 5 | 2 | Only physically (present?) | 37.5 |
| S | Better speaking | 1 | 1 | 1 | 2 | 3 | 4 | 5 | 3 | Poorer speaking | 75 |
| L | Better listening | 1 | 1 | 2 | 1 | 4 | 4 | 4 | 3 | Poorer listening | 75 |
| O | Better overall | 1 | 1 | 1 | 1 | 3 | 4 | 4 | 3 | Poorer overall | 75 |

clear that this concept disregarded the language (English or Japanese) the chatting was done in. He felt it may, however, impact the concept of speaking to some degree.

Ryo was reported as having ability but not participating in the class. He was a leader, but the participant reported that he did it "in a negative way." Although Ryo often talked to others, perhaps distracting those of lower ability, he was easily able to answer questions. On the other hand, weaker students, such as Osamu or Akiko, were less focused on the class and less able to handle the material from the outset. It may be that these students would have been better placed in a different level of class, more in line with their ability to handle higher or lower material. The extremes of ability were judged by the participant to potentially impact the performance of others class members on groupwork tasks, which would include the panel discussion for the main assessment.

On a qualitative level, the teacher pointed out that other non-academic factors impacted the grade. For example, the 9 o'clock start was an issue, inasmuch as he described the class as "not first-period people." In particular, the participant pointed out that the scores on elicited constructs show that Yukiko's performance and grade may not match. Her performance was described as similar to Chieko's. Yukiko had performed reasonably, but had fallen foul of an attendance and lateness policy. The participant reported that Yukiko had taken steps to overcome in the next semester.

More quantitative analysis offers further insight. Figure 1 shows a dendrogram of elements and constructs. This diagram gives a pictorial representation of how constructs may correlate. Koh and Masako (near the bottom of Figure 1), for example, have very similar scores on constructs, and are joined by short lines at the bottom of the figure. Similarly, constructs C03 and C04 (near the top) score exactly

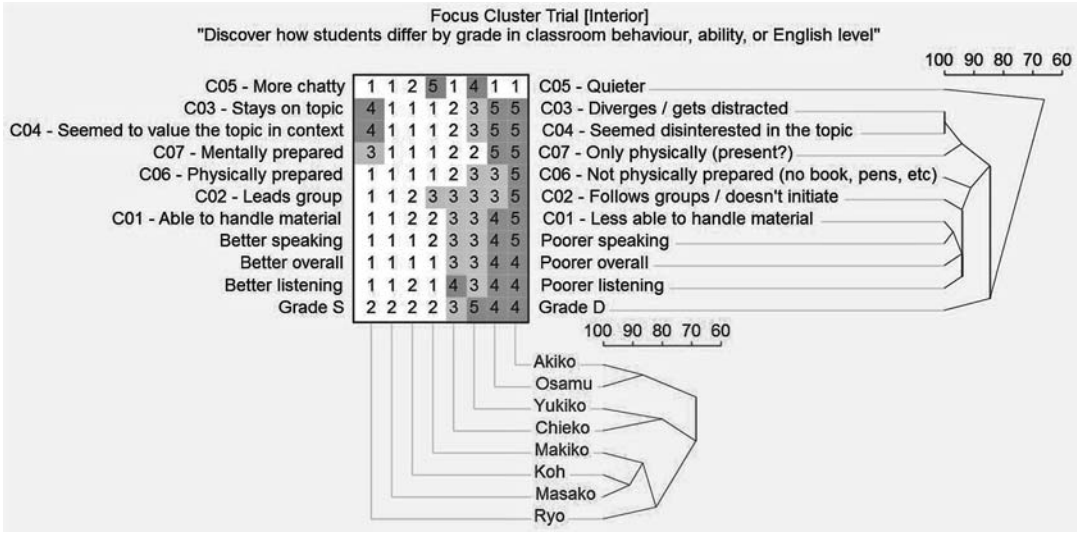


**Figure 1 A Dendrogram of the Data Provided by the Participant**

the same for all elements, and are connected by an almost vertcal line (top right), suggesting they may be describing something similar.

The dendrogram confirms that Yukiko and Chieko are more similar to each other than to others, despite the difference in their grade. This confirms the information supplied when the teacher described the issues of attendance, as detailed in the qualitative data.

The supplied English ability constructs all cluster closely on Figure 1, along with C01 (more/less able to handle material). Such a correlation is a positive sign for the class, and a lack of correspondence may represent a threat to the validity the grade to some extent. Additionally, however, these constructs also seem related to C02 and C06, forming a kind of cluster or branch of constructs. Constructs C03, C04, and C07 seem to form a second branch on Figure 1 (near the top). These two branches seem to combine before joining the grade construct. In contrast, C05 joins after the grade, in relative isolation. This may suggest that the grade is less affected by how gregarious a student is, and more affected by the other components mentioned.

Principal Component Analysis (Figure 2) revealed four factors, two of which account for 88.9% of variance. This is close to (but still under) the 90% threshold recommended in Jankowicz (2004). The two components derived from the data are displayed as axes in the figure.

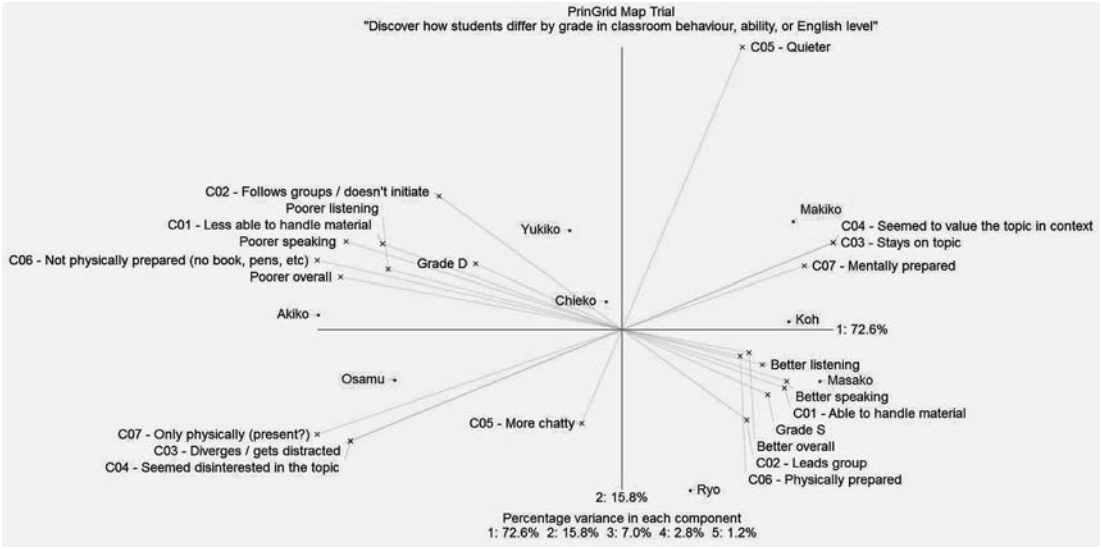


**Figure 2 Principal Component Analysis of Data from the Participant**

The grade line seems to follow C01 (handling material) relatively closely. The supplied ability constructs, along with C01 and C06, form a common thread that moves from the top left to bottom right of the graph, closer to the x axis than the y axis. A second thread of constructs seems to track through the opposite quadrants, with C03, C04, and C07 almost as a mirror image of the grade-line thread. Once again, construct C05 does not track with other threads, suggesting it may be more peripheral.

Students described by the participant are displayed at various single points on the graph. Their proximity to constructs may reflect the impact of those constructs, particularly where they group closely together. Masako, for example, can be found at a point near the S end of the grade trajectory, making her a possible exemplar of the grade. Ryo, the other student to receive the same grade as Masako, has been pulled away from this trajectory, suggesting that he does not fit the normal grade pattern. Osamu and Makiko, however, seem to be close to the second branch of components, appearing at opposite ends. They may be more strongly affected by this component. Koh and Akiko appear at opposite ends of the x axis, closely following the main component extracted from the principal component analysis.

## V. Discussion

The class under investigation is part of a broader English program as part of a general education program. As such, there is no minimum level of English required, and students are accepted as is, reflecting different prior learning experiences.

Accounting for this is problematic (American Educational Research Association et al., 2014), and using a set of benchmarks or common standards would create a highly inequitable situation.

Teachers must therefore work with a textbook and find a way to create a grade. The choices of approaches to grading are reasonably open in this setting. While this teacher used performance-based assessments with presentations and discussions, other teachers used listening tests generated from the textbook, or discrete language items covering vocabulary or grammar. Nonetheless, all students received a grade called *English A* on their transcript. This reflects a more developmental than standards-based approach. However, the information carried by the grade becomes unclear when proficiency and development may be conflated.

Proficiency clearly forms a stronger part of the grade than development, as evidenced by construct C01, and exemplified by Masako's performance. Indeed, Ryo was reported as scoring highly, even though he was distracting others. On the other hand, Osamu and Akiko seem to have less ability to handle material from the outset. Given that all students in the class scored similarly on the placement instrument, this seems to represent a considerable degree of error to account for. Although placement systems are often thought of as being accurate, any test has a standard error of measurement. Teachers therefore need strategies for dealing fairly with such disparities in ability.

It seems the teacher was also looking for something akin a positive academic outlook, perhaps close to that found in other work on classroom-based assessment. Sun and Cheng, for example, found that teachers in China rewarded skills that would enable students in their academic career (Sun & Cheng, 2014). This may be what is being shown in concepts such as valuing the topic (C04), or being physically (C06) or mentally (C07) prepared. From this perspective, it may be surprising that being "mentally prepared" had the relatively low quantitative impact observed here. A broader sample of classes and teachers may help to establish this in more concrete terms, and allow it to be more concretely utilised in the grading process and the interpretation of grades by nonclassroom-based stakeholders.

Both Makiko and Koh originally received the same letter grade as Ryo and Masako, and both score well on the poles of constructs associated with higher grades. However, because their numeric score was slightly lower, the teacher was forced to downgrade them as a result of the grading policy. Error in assessment scores is a feature of testing, and may be commonly reported for large-scale tests, such as IELTS or TOEIC. Given the improvised and pre-fabricated nature of much classroom-based assessment, it may be incumbent on teachers to seek further evidence that a particular score is both fair and appropriate for a particular student.

It is worth nothing, however, that it is the school, rather than the teacher, that

is requiring a score. While class scores are often submitted as graded out of 100, it should be remembered that this is not a percentage. The construct of the course, which is vague to begin with, is unlikely to simply divide into 100 equal pieces. The ultimate grades received are simply a nominal scale, and should be treated as descriptions rather than mathematical entities. The numeric scores from teachers do not offer a ratio scale, and should not be taken a perfect measurement. Requiring such a score may be placing the teacher in the way of more conflict between the role of coach and judge (Bishop, 1992).

In particular, the developmental aspect of the course may need more concrete operationalization to be more equitable. Homework here is graded as "complete/ incomplete," and those receiving a lower grade generally had issues with attendance or engagement. The participant's "true scores" may have more heavily reflected the second strand of the grade construct observed in the numeric and qualitative data, but may also be more problematic in terms of useful measurement.

With regard to the use of repertory grid technique in this setting, it was found to be a useful way to structure an interview, as was suggested by Hadley (2017). In common with many qualitative approaches, it helps to view the situation through the eyes of the participant, in a way that is congruent with Grounded Theory (Bryant, 2017) or with Thematic Analysis (Guest et al., 2011). Particularly useful, however, was the collaborative approach it involved, making the interviewer and interviewee partners in the research process, consistent with exploratory practice (Hanks, 2017). While the perspective of the teacher in this project takes precedence (Jankowicz, 2004), the process of negotiating what the teacher means and communicating it in terms of the scales opens up the setting to dialogue and exploration in a way that would simply not be possible in a more quantitative way. The numeric data provides a way in which to begin to interpret and build upon the qualitative data, and vice versa (Bazeley, 2018).

As with any technique, there are limitations on what repertory grids can and cannot do. Given that a single interview took 80 minutes, RGT has considerable issues relating to scale, making it too time-consuming to be broadly used. Much of the time is taken negotiating the poles, so skilful interviewing techniques are required. In addition, although RGT provides some rich data connected to the construct, it is hard to find causality. Are students getting high scores because they value the topic, or do they value the topic because of the high scores? The reality is probably a little of both, but strong evidence either way is unlikely to come from this method.

## Theoretical Implications

Despite, the above weaknesses, the process of creating, rejecting, and refining

the constructs using RGT offers a reflective tool for all involved in the research process. Although Personal Construct Theory emphasizes the individual's perceptions, the process may help in identifying context specific changes or good practice. The convergence and divergence of qualitative and quantitative data offers more depth, such that we can see why two members of similar ability get different grades (Yukiko and Chieko). In some ways, issues such as working with members of the same gender or the 9 o'clock start may not admit to the quantitative data approach, but can be logged as issues requiring further attention.

The emphasis on starting with teachers who know the history of their specific learners provides powerful comparison of desired assessment targets and observed successes or failures. In addition, the examination of the constructs in a collaborative way means that the teacher also gains through the interview process. As such, repertory grids may have a place both as a research tool and as a tool for broader faculty and institutional development. This technique may have implications for research beyond assessment, such as teacher beliefs or isolating teaching practice that aids employability of students.

## VI. Conclusions

Although this was a sample of one individual from a team of 80 teachers, repertory grid technique shows some merit in this setting. The grade given in *English A* by the participant seemed to have a strong element of proficiency, but also carried a performance element, showing how students have used academic enablers and developed through the class. As described by this participant, it may be that the grade shows the application of proficiency rather than proficiency itself. This is a distinction that may be of value to other stakeholders. Further application of the technique, combined with other approaches, may help to give a clearer idea of what stakeholders may reasonably infer from the grade given. The present label of "English" is perhaps too vague for students, parents, or potential employers to use meaningfully.

The absence of benchmarks for EFL assessment in this setting is a practical necessity, however. Teachers are working hard to provide quality language education, and techniques such as this may also help to highlight teacher achievements. Particularly given the requirement for a teacher to adjust to the students in front of them, this research method may contribute to efforts to improve the offerings of language teachers, by offering guidance to teachers based on their colleagues' practice, and demonstrating what a good grade may represent.

## References

Allwright, D. (2005). Developing principles for practitioner research: The case of exploratory practice. *The Modern Language Journal, 89,* 353-366. https://doi.org/10.1111/j.1540-4781.2005.00310.x

American Educational Research Association, American Psychological Association, & National Council on Research in Education. (2014). *Standards for educational and psychological testing.* American Educational Research Association.

Bazeley, P. (2013). *Qualitative data analysis.* SAGE Publications, Ltd.

Bazeley, P. (2018). *Integrating analyses in mixed methods research.* SAGE Publications Ltd.

Bishop, J. (1992). Why US students need incentives to learn. *Educational Leadership, 49,* 15-18.

Block, D. (1997). Learning by listening to language learners. *System, 25*(3), 347-360. https://doi.org/10.1016/S0346-251X(97)00027-4

Borg, S. (2015). *Teacher cognition and language education: Research and practice.* Bloomsbury Academic. https://www.amazon.co.jp/gp/product/B00T4AL1KK/ref=ppx_yo_dt_b_d_asin_title_o03?ie=UTF8&psc=1

Brookhart, S. M., & McMillan, J. H. (2019). *Classroom Assessment and Educational Measurement.* Routledge.

Brown, G. T. L., & Hirschfeld, G. H. F. (2008). Students' conceptions of assessment: Links to outcomes. *Assessment in Education: Principles, Policy & Practice, 15*(1), 3-17. https://doi.org/10.1080/09695940701876003

Brown, G. T. L., & Wang, Z. (2013). Illustrating assessment: How Hong Kong university students conceive of the purposes of assessment. *Studies in Higher Education, 38,* 1037-1057. https://doi.org/10.1080/03075079.2011.616955

Bryant, A. (2017). *Grounded Theory and Grounded Theorizing: Pragmatism in Research Practice.* Oxford University Press.

Cheng, L., & Fox, J. (2017). *Assessment in the language classroom: Teachers supporting student learning.* Red Globe Press.

Coombe, C., Vafadar, H., & Mohebbi, H. (2020). Language assessment literacy: What do we need to learn, unlearn, and relearn? *Language Testing in Asia, 10*(1), 3. https://doi.org/10.1186/s40468-020-00101-6

Creswell. (2014). *A concise introduction to Mixed Methods Research.* SAGE Publications, Ltd.

Davidson, F., & Lynch, B. K. (2002). *Testcraft: A teacher's guide to writing and using language test specifications.* Yale University Press.

Denicolo, P., Long, T., & Bradley-Cole, K. (2016). *Constructivist approaches and research methods: A practical guide to exploring personal meanings.* SAGE Publications Ltd.

Fulcher, G. (2010). *Practical language testing.* Routledge.

Gipps, C. (1994). *Beyond testing.* Routledge.

Guest, G., MacQueen, K. M., & Namey, E. E. (2011). *Applied thematic analysis.* SAGE Publications, Ltd.

Hadley, G. (2015). *English for Academic Purposes in neoliberal universities: A critical grounded theory.* Springer.

Hadley, G. (2017). *Grounded Theory in applied linguistics research: A practical guide.* Routledge.

Hanks, J. (2017). *Exploratory practice in language teaching: Puzzling about principles and practices.* Palgrave Macmillan.

Harding, L., & Kremmel, B. (2016). Teacher assessment literacy and professional development. In D. Tsagari & J. Banerjee, *Handbook of second language assessment.* (Vol.12, pp.413-427). Mouton de Gruyter.

Heritage, M., & Harrison, C. (2020). *The power of assessment for learning: Twenty years of research and practice in UK and US classrooms.* Corwin.

Hill, K. (2017). Understanding classroom-based assessment practices: A precondition for teacher assessment literacy. *Papers in Language Testing and Assessment, 6,* 1-17.

Honey, P. (1979). The repertory grid in action: How to use it to conduct an attitude survey. *Industrial and Commercial Training, 11,* 452-459. https://doi.org/10.1108/eb003756

ILTA. (2007). *Guidelines for practice.* ILTA Guidelines for Practice. http://www.iltaonline.com/index.php/enUS/resources/ilta-guidelines-for-practice

Jankowicz, D. (2004). *The easy guide to Repertory Grids.* Wiley.

Kelly, G. (1992). *The Psychology of Personal Constructs: Volume One: Theory and Personality.* Routledge.

Lantolf, J. P., & Poehner, M. E. (2014). *Sociocultural theory and the pedagogical imperative in L2 education: Vygotskian praxis and the research/practice divide.* Routledge.

Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis: A methods sourcebook. 3rd.* SAGE Publications, Ltd.

Smith, J. K. (2003). Reconsidering reliability in classroom assessment and grading. *Educational Measurement: Issues and Practice, 22,* 26-33. https://doi.org/10.1111/j.1745-3992.2003.tb00141.x

Sun, Y., & Cheng, L. (2014). Teachers' grading practices: Meaning and values assigned. *Assessment in Education: Principles, Policy & Practice, 21,* 326-343. https://doi.org/10.1080/0969594X.2013.768207

Swain, P. M., Kinnear, D. P., & Steinman, L. (2015). *Sociocultural Theory in second language education: An introduction through narratives* (2nd ed.). Multilingual Matters.

Turner, C. E. (2012). Classroom assessment. In G. Fulcher & F. Davidson, *The Routledge Handbook of Language Testing* (pp.64-78). Routledge. https://doi.org/10.4324/9780203181287.ch4

*WebGrid Plus.* (2017). http://webgrid.uvic.ca/

**Appendix 1 - Modified interview record sheet**